IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Utility Patent Application

ROBUST BAYESIAN MIXTURE MODELING

Inventor(s):
Christopher M. Bishop
Markus Svensén

CLIENT'S DOCKET NO. MS305414.01
ATTORNEY'S DOCKET NO. MS1-1673US

EV395541585

ROBUST BAYESIAN MIXTURE MODELING

Technical Field

The invention relates generally to statistical analysis and machine learning algorithms, and more particularly to robust Bayesian mixture modeling.

Background

Mixture models are common tools of statistical analysis and machine learning. For example, when trying to model a statistical data distribution, a single Gaussian model may not adequately approximate the data, particularly when the data has multiple modes or clusters (e.g., has more than one peak).

As such, a common approach is to use a mixture of two or more Gaussian components, fitted with a maximum likelihood, to model such data. Nevertheless, even a mixture of Gaussians (MOG) presents modeling problems, such as inadequate modeling of outliers and severe overfitting. For example, there are singularities in the likelihood function arising from the collapse of components onto individual data points – a pathological result.

Some problems with a pure MOG can be elegantly addressed by adopting a Bayesian framework to marginalize over the model parameters with respect to appropriate priors. The resulting Bayesian model likelihood can then be maximized with respect to the number of Gaussian components in the mixture, if the goal is model selection, or combined with a prior over the number of the components, if the goal is model averaging. One benefit to a Bayesian approach using a mixture of Gaussians is the elimination of maximum likelihood singularities, although it still lacks robustness to outliers. In addition, in the

Bayesian model selection context, the presence of outliers or other departures from the empirical distribution of Gaussianity can lead to errors in the determination of the number of clusters in the data.

Summary

Implementations described and claimed herein address the foregoing problems using a Bayesian treatment of mixture models based on individual components having Student distributions, which have heavier tails compared to the exponentially decaying tails of Gaussians. The mixture of Student distribution components is characterized by a set of modeling parameters. Tractable approximations of the posterior distributions of individual modeling parameters are optimized and used to generate a data model for a set of input data.

In some implementations, articles of manufacture are provided as computer program products. One implementation of a computer program product provides a computer program storage medium readable by a computer system and encoding a computer program. Another implementation of a computer program product may be provided in a computer data signal embodied in a carrier wave by a computing system and encoding the computer program.

The computer program product encodes a computer program for executing a computer process on a computer system. A modeling parameter is selected from a plurality of modeling parameters characterizing a mixture of Student distribution components. A tractable approximation of a posterior distribution for the selected modeling parameter is computed based on an input set of data and a current estimate of a posterior distribution of at least one unselected modeling parameter in the plurality of modeling parameters. A lower bound of a log marginal

likelihood is computed as a function of current estimates of the posterior distributions of the modeling parameters. The current estimates of the posterior distributions of the modeling parameters include the computed tractable approximation of the posterior distribution of the selected modeling parameter. A probability density that models the input set of data is generated, if the lower bound is satisfactorily optimized. The probability density includes the mixture of Student distribution components, which is characterized by the current estimates of the posterior distributions of the modeling parameters.

In another implementation, a method is provided. A modeling parameter is selected from a plurality of modeling parameters characterizing a mixture of Student distribution components. A tractable approximation of a posterior distribution for the selected modeling parameter is computed based on an input set of data and a current estimate of a posterior distribution of at least one unselected modeling parameter in the plurality of modeling parameters. A lower bound of a log marginal likelihood is computed as a function of current estimates of the posterior distributions of the modeling parameters. The current estimates of the posterior distributions of the modeling parameters include the computed tractable approximation of the posterior distribution of the selected modeling parameter. A probability density that models the input set of data is generated, if the lower bound is satisfactorily optimized. The probability density includes the mixture of Student distribution components, which is characterized by the current estimates of the posterior distributions of the modeling parameters.

In another implementation, a system is provided. A tractable approximation module computes a tractable approximation of a posterior

distribution for the selected modeling parameter based on an input set of data and a current estimate of a posterior distribution of at least one unselected modeling parameter in the plurality of modeling parameters. A lower bound optimizer module computes a lower bound of a log marginal likelihood as a function of current estimates of the posterior distributions of the modeling parameters. The current estimates of the posterior distributions of the modeling parameters include the computed tractable approximation of the posterior distribution of the selected modeling parameter. A data model generator generates a probability density modeling the input set of data, if the lower bound is satisfactorily optimized. The probability density includes the mixture of Student distribution components. The mixture of Student distribution components is characterized by the current estimates of the posterior distributions of the modeling parameters.

Other implementations are also described and recited herein.

Brief Descriptions of the Drawings

FIG. 1 illustrates exemplary probability distributions for modeling a data set.

FIG. 2 illustrates exemplary operations for robust Bayesian mixture modeling.

FIG. 3 illustrates an exemplary robust Bayesian mixture modeling system.

FIG. 4 illustrates a system useful for implementing an embodiment of the present invention.

Detailed Description

FIG. 1 illustrates exemplary probability distributions 100 for modeling a data set. A single Gaussian distribution 102 models an input data set of independent identically distributed (idd) data 104. Note that the mean 106 of the single Gaussian distribution 102 is pulled substantially to the right in order accommodate the outlier data element 106, thereby compromising the accuracy of the Gaussian model as it applies to the given data set 104. In addition, the standard deviation of the distribution 102 is undesirably increased by the outlier 106.

In order to improve the modeling of the data 104, a mixture of Gaussian distributions 108 may be used. However, fitting the mixture 108 to the data set 104 using a maximum likelihood approach does not yield a usable optimal number of components because the maximum likelihood approach favors an ever more complex model, leading to the undesirable extreme of individual, infinite magnitude Gaussian distribution component for individual data point. While overfitting of Gaussian mixture models can be addressed to some extent using Bayesian inference, even then, Gaussian mixture models continue to lack robustness as to outliers.

A mixture of Student distributions 110 can demonstrate a significant improvement in robustness as compared to a mixture of Gaussian distributions. However, there is no closed form solution for maximizing the likelihood under a Student distribution. Furthermore, the maximum likelihood approach does not address the problem of overfitting. Therefore, a mixture of Student distributions 110 combined with a tractable Bayesian treatment to fit the Student

mixture to the input data 104 addresses these issues, as illustrated in FIG. 1. However, no satisfactory method or system for obtaining a tractable Bayesian treatment of Student mixture distributions has previously been demonstrated. As such, in one implementation, robust Bayesian mixture modeling obtains a tractable Bayesian treatment of Student mixture distributions based on variational inference. In another implementation, a tractable approximation may be obtained using Monte Carlo-based techniques.

Robust Bayesian mixture modeling is based on a mixture of component distributions given by a multivariate Student distribution, also known as a *t*-distribution. A Student distribution represents a generalization of a Gaussian distribution and, in the limit $\nu \to \infty$, the Student distribution reduces to a Gaussian distribution with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$ (i.e., inverse covariance). For finite values of ν, the Student distribution has heavier tails than the corresponding Gaussian having the same $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$.

A Student distribution over a *d*-dimensional random variable $\mathbf{x}$ may be represented in the following form:

$$S\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu\right) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{d}{2}}}\left(\frac{\Delta^2}{\nu}+1\right)^{-\frac{\nu+d}{2}} \qquad (1)$$

where $\Delta^2 = (x-\mu)^{\mathsf{T}}\Lambda(x-\mu)$ represents the squared Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$.

In contrast to the Gaussian distribution, no closed form solution for maximizing likelihood exists under a Student distribution. However, the Student distribution may be represented as an infinite mixture of scaled Gaussian

distributions over **x** with an additional random variable u, which acts as a scaling parameter of the precision matrix $\mathbf{\Lambda}$, such that the Student distribution may be represented in the following form:

$$S\left(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Lambda},\nu\right)=\int_0^\infty N\ \left(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Lambda}u\right)G\left(u|\frac{\nu}{2},\frac{\nu}{2}\right)du \tag{2}$$

where $N\left(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Lambda}\right)$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{\Lambda}u$, and $G\left(u|a,b\right)$ represents the Gamma distribution. For each observation of **x** (i.e., of N observations), a corresponding implicit posterior distribution over the variable u exists.

The probability density of mixtures of M Student distributions may be represented in the form:

$$p\left(\mathbf{x}|\left\{\boldsymbol{\mu}_m,\mathbf{\Lambda}_m,\nu_m\right\},\pi\right)=\sum_{m=1}^{M}\pi_m S\ \left(\mathbf{x}|\boldsymbol{\mu}_m,\mathbf{\Lambda}_m,\nu_m\right) \tag{3}$$

where the mixing coefficients $\pi=\left(\pi_1,...,\pi_M\right)^{\mathrm{T}}$ satisfy $0\le\pi_m\le 1$ and $\sum_{m=1}^{M}\pi_m=1$.

In order to find a tractable treatment of this model, the mixture density of Equation (3) may be expressed in terms of a marginalization over a binary latent labeling variable **s** of dimensions $N\times M$ (i.e., N representing the number of data elements and M representing the number of Student distribution components in the mixture) and the unobserved variable u_{nm}, also of dimensions $N\times M$ when applied to a mixture. Variable **s** has components $\{s_{nj}\}$, such that $s_{nm}=1$ and $s_{nj}=0$ for $j\neq m$, resulting in:

$$p\left(\mathbf{x}_n|\mathbf{s},\left\{\mu_m,\Lambda_m,\nu_m\right\}\right)=\prod_{n,m}^{N,M}S\ \left(\mathbf{x}|\mu_m,\Lambda_m,\nu_m\right)^{s_{nm}} \tag{4}$$

with a corresponding prior distribution over **s** of the form:

$$p(\mathbf{s}|\pi) = \prod_{n,m}^{N,M} \pi_m^{s_{nm}} \tag{5}$$

It can be verified that marginalization of the product of Equations (4) and (5) over the latent variable **s** recovers the Student distribution mixture of Equation (3).

An input data set **X** includes N idd observations $\mathbf{x}_n$, where n=1,...,N, which are assumed to be drawn independently from the distribution characterized by Equation (3). Thus, for each data observation $\mathbf{x}_n$, a corresponding discrete latent variable $\mathbf{s}_n$ specifies which component of the mixture generated that data point, and continuous latent variable u_{nm} specifies the scaling of the precision for the corresponding equivalent Gaussian distribution from which the data was hypothetically generated.

In addition to the prior distribution over **s**, prior distributions for the modeling parameters $\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m$, and $\boldsymbol{\pi}$, are used in a Bayesian treatment of probability density estimation. As such, distributions of the modeling parameters are used rather than the parameters themselves. In one implementation, for tractability, conjugate priors from the exponential family have been chosen in the form:

$$p(\boldsymbol{\mu}_m) = N\ (\boldsymbol{\mu}_m|\mathbf{m}, \rho\mathbf{I}) \tag{6}$$

$$p(\boldsymbol{\Lambda}_m) = W\ (\boldsymbol{\Lambda}_m|\mathbf{W}_0, \eta_0) \tag{7}$$

$$p(\pi) = D\ (\pi|\boldsymbol{\alpha}) \tag{8}$$

wherein $W(\Lambda|\square,\square)$ represents the Wishart distribution and $D(\pi|\square)$ represents the Dirichlet distribution. The prior p(u) is implicitly defined in Equation (2) to equal the Gamma distribution $G\left(u|\frac{\nu}{2},\frac{\nu}{2}\right)$.

It should be understood that prior distributions may be selected from other members of the exponential family in alternative embodiments. The parameters of the prior distributions on $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are chosen to give broad distributions (e.g., in one implementation, $\mathbf{m}_0=0$, $\rho_0=10^{-3}$, $\mathbf{W}_0=\mathbf{I}$, $\eta_0=1$. For the prior distribution over π, $\alpha=\{\alpha_m\}$ are interpreted as effective numbers of prior observations, with $\alpha_m=10^{-3}$.

Exact inference of the Bayesian model is intractable. However, with the choice of exponential distributions to represent the prior distributions of the modeling parameters, tractable approximations are possible. In one implementation, for example, a tractable approximation may be obtained through Monte Carlo techniques.

In another implementation, variational inference may be employed to obtain tractable approximations of the posterior distributions over the identified stochastic modeling parameters, which in one implementation includes $\{\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m\}$, π, and $\{\mathbf{s}_m,\mathbf{u}_n\}$. (Another modeling parameter, v, is treated in a deterministic (i.e., non-stochastic) fashion; however, only one such parameter exists per mixture component.).

In variational inference, the log-marginal likelihood is maximized. One form of the log-marginal likelihood is shown:

$$\ln\prod_n^N p(\mathbf{x}_n|m_0,\rho_0,W_0,\eta_0)= \ln\int\prod_n^N p(\mathbf{x}_n,u_n|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu)p(\boldsymbol{\mu}|m_0,\rho_0)\,p(\boldsymbol{\Lambda}|\mathbf{W},\eta)\,du_n d\boldsymbol{\mu} d\boldsymbol{\Lambda} \quad (9)$$

This quantity cannot be maximized directly. However, Equation (9) can be re-written as follows:

$$\ln \int p(\mathbf{X}|\theta)\, p(\theta|m_0,\rho_0,\mathbf{W}_0,\eta_0,\nu)d\theta = \int q(\theta)\ln\frac{p(\mathbf{X},\theta|m_0,\rho_0,\mathbf{W}_0,\eta_0,\nu)}{q(\theta)}d\theta - \int q(\theta)\ln\frac{p(\theta,\mathbf{X}|m_0,\rho_0,\mathbf{W}_0,\eta_0,\nu)}{q(\theta)}d\theta \quad (10)$$

where $\mathbf{X}=\{\mathbf{x}_n\}$, $\theta=\{\boldsymbol{\mu},\boldsymbol{\Lambda},\mathbf{u}\}$, $\mathbf{u}=\{u_n\}$, and $q(\theta)$ is the so-called variational distribution over $\boldsymbol{\mu},\boldsymbol{\Lambda}$, and $\mathbf{u}$, such that $q(\theta)=q(\boldsymbol{\mu})q(\boldsymbol{\Lambda})q(\mathbf{u})$ (assuming $q(\boldsymbol{\mu})$, $q(\boldsymbol{\Lambda})$, and $q(\mathbf{u})$ are independent).

The second term of Equation (10) is the Kullback-Leibler (KL) divergence between $q(\theta)$ and $p(\theta|\{x_n\},m_0,\rho_0,\mathbf{W}_0,\eta_0,\nu)$, which is non-negative and zero only if the two distributions are identical. Thus, the first term can be understood as the lower bound of the log-marginal likelihood $\Lambda(q)$. Therefore, seeking to minimize the second term of Equation (10) amounts to maximizing the lower bound $\Lambda(q)$.

Accordingly, one way to represent the lower bound $\Lambda(q)$ is shown:

$$L(q) \equiv \int q(\theta)\ln\left\{\frac{p(\mathbf{X},\theta)}{q(\theta)}\right\}d\theta \le \ln p(\mathbf{X}) \quad (11)$$

where θ represents the set of all unobserved stochastic variables.

In Equation (11), $q(\theta)$ represents the variational posterior distribution, and $p(\mathbf{X},\theta)$ is the joint distribution over the stochastic modeling parameters. The difference between the right hand side of Equation (11) and $\Lambda(q)$ is given by the

KL divergence KL($q||p$) between the variational posterior distribution $q(\theta)$ and the true posterior distribution $p(\theta,\mathbf{X})$.

Given the priors of Equations (5), (6), (7), and (8), the variational posterior distributions $q(\cdot)$ for $\mathbf{s}$, π, $\boldsymbol{\mu}_m$, $\boldsymbol{\Lambda}_m$, and u may be computed.

For $q(\mathbf{s})$, where $\mathbf{s}$ represents the labeling parameters:

$$q(\mathbf{s}) = \prod_{n,m}^{N,M} p_{nm}^{s_{nm}} \tag{12}$$

where

$$p_{nm} = \frac{r_{nm}}{\sum_{m=1}^{M} r_{nm'}} \tag{13}$$

where, in turn,

$$r_{nm} = \exp\left(\langle \ln \pi_m \rangle + \frac{1}{2} \langle \ln |\boldsymbol{\Lambda}_m| \rangle + \frac{d}{2} \langle \ln u_{nm} \rangle - \frac{\langle u_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2} - \frac{d}{2} \ln 2\pi \right) \tag{14}$$

Although the last term in the argument for the exponential cancels out in Equation (13). In addition,

$$\langle \ln |\boldsymbol{\Lambda}_m| \rangle = d \ln 2 - \ln |\mathbf{W}| + \sum_{i=1}^{d} \Psi\left(\frac{\eta + 1 - i}{2} \right) \tag{15}$$

$$\langle \Delta_n^2 \rangle = \mathbf{x}_n^{\mathsf{T}} \eta_m \mathbf{W}_m \mathbf{x}_n - 2\mathbf{x}_n^{\mathsf{T}} \eta_m \mathbf{W}_m \mathbf{m}_m + \mathrm{Tr}\left[\left(\mathbf{m}_m \mathbf{m}_m^{\mathsf{T}} + R_m^{-1} \right) \eta_m \mathbf{W}_m \right] \tag{16}$$

and

$$\langle s_{nm} \rangle = p_{nm} \tag{17}$$

For $q(\pi)$, where π represents the mixing coefficients:

$$q(\pi) = D(\pi|\alpha) \tag{18}$$

where

$$\alpha_m = \sum_{n=1}^{N} \langle s_{nm} \rangle + \hat{\alpha}_m \tag{19}$$

and

$$\langle \pi_m \rangle = \frac{\alpha_m}{\alpha_0} \tag{20}$$

where $\alpha_0 = \sum_{m'} \alpha_{m'}$ and m'=1,...,M. Furthermore, $\langle \ln \pi_m \rangle = \Psi(\alpha_m) - \Psi(\alpha_0)$, where

$$\Psi(a) = \frac{d \ln \Gamma(a)}{da} \tag{21}$$

For $q(\mu_m)$, where μ_m represents the mean of the m^{th} Student distribution component in the mixture:

$$q(\boldsymbol{\mu}_m) = N(\boldsymbol{\mu}_m | m_m, \mathbf{R}_m) \tag{22}$$

where

$$\mathbf{R}_m = \langle \Lambda_m \rangle \sum_{n=1}^{N} \langle w_{nm} \rangle + \rho_0 \mathbf{I} \tag{23}$$

$$m_m = \mathbf{R}_m^{-1} \left(\langle \boldsymbol{\Lambda}_m \rangle \sum_{n=1}^{N} \langle w_{nm} \rangle \mathbf{x}_n + \rho_0 m_0 \right)$$

and

$$\langle w_{nm} \rangle = \langle s_{nm} \rangle \langle u_{nm} \rangle \tag{24}$$

For $q(\mathbf{\Lambda}_m)$, where $\mathbf{\Lambda}_m$ represents the precision matrix of the m^{th} Student distribution component in the mixture:

$$q(\mathbf{\Lambda}_m) = W\left(\mathbf{\Lambda}_m \middle|_m \mathbf{W}_m, \eta_m\right) \tag{25}$$

where

$$\mathbf{W}_m^{-1} = \mathbf{W}_0^{-1} + \sum_n^N \langle w_{nm} \rangle \left(\mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} - \mathbf{x}_n m_m^{\mathsf{T}} - m_m \mathbf{x}_n^{\mathsf{T}} + \left(m_m m_m^{\mathsf{T}} + R_m^{-1} \right) \right) \tag{26}$$

and

$$\eta_m = \eta_0 + \hat{s}_m \tag{27}$$

where $\hat{s}_m = \sum_n^N \langle s_{nm} \rangle$.

For $q(\mathbf{u})$, where $\mathbf{u}$ represents the scaling parameters of the precision matrices:

$$q(u_{nm}) = G\left(u_{nm} \middle| a_{nm}, b_{nm}\right) \tag{28}$$

where

$$a_{nm} = \frac{\nu_m + \langle s_{nm} \rangle d}{2} \tag{29}$$

where d represents the dimensionality of the data,

$$b_{nm} = \frac{\nu_m + \langle s_{nm} \rangle \langle \Delta_{nm}^2 \rangle}{2} \tag{30}$$

and

$$\left\langle \Delta_{nm}^2 \right\rangle = \mathbf{x}_n^{\mathsf{T}} \eta_m \mathbf{W}_m \mathbf{x}_n - 2\mathbf{x}_n^{\mathsf{T}} \eta_m \mathbf{W}_m m_m + \mathrm{Tr}\left[\left(m_m m_m^{\mathsf{T}} + R_m^{-1} \right) \eta_m \mathbf{W}_m \right] \tag{31}$$

A constrained family of distributions for $q(\theta)$ is chosen such that the lower bound $\Lambda(q)$ becomes tractable. The optimal member of the family can then be determined by maximization of $\Lambda(q)$, which is equivalent to minimization of the KL divergence. Thus, the resulting optimal solution for $q(\theta)$ represents an approximation of the true posterior of $p\left(\theta \middle| \{x_n\}, \mathbf{m}_0, \rho_0, \mathbf{W}_0, \eta_0, \nu\right)$, assuming a factorized variational distribution for $q(\theta)$ of:

$$q(\theta) = q\left(\{\boldsymbol{\mu}_m\}\right) q\left(\{\boldsymbol{\Lambda}_m\}\right) q(\pi) q\left(\{\mathbf{s}_n\}\right) q\left(\{\mathbf{u}_n\}\right) \tag{32}$$

A free-form variational optimization is now possible with respect to each of the individual variational factors of Equation (32). Because the variational factors are coupled, the variational approximations of the factors are computed iteratively by first initializing the distributions, and then cycling to each factor in turn and replacing its current estimate by its optimal solution, given the current estimates for the other factors, to give a new approximation of $q(\theta)$. Interleaved with the optimization with respect to each of the individual variational factors, the lower bound is optimized with respect to each of the non-stochastic parameters v_m by employing standard non-linear optimization techniques. The lower bound $\Lambda(q)$ is then computed using the new approximation of $q(\theta)$ for the current iteration

In one implementation, the iteration continues until the lower bound $\Lambda(q)$ changes by less than a given threshold. In an alternative implementation, $q(\theta)$ may also be tested prior to computation of the lower bound $\Lambda(q)$ in each iteration,

such that if the value of $q(\theta)$ changes by less than another given threshold, then the iteration skips the computation and testing of the lower bound $\Lambda(q)$ and exits the loop. In yet another implementation, individual factors of Equation (32) may be tested to determine whether to terminate the optimization of the modeling parameters.

In the described approach, approximate posterior distributions of the stochastic modeling parameters $\{\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m\}$, π, and $\{\mathbf{s}_m, \mathbf{u}_n\}$, as well as a value of the modeling parameter v, are determined. Given these modeling parameters, the Student mixture density of Equation (3) can be obtained to model the input data.

FIG. 2 illustrates exemplary operations 200 for robust Bayesian mixture modeling. A receiving operation 202 receives prior distributions of each modeling parameter in the set of modeling parameters for a mixture of Student distributions. In one implementation, the prior distributions may be computed using the Equations (5), (6), (7), and (8), although other prior distributions may be used in alternative embodiments. As such, an operation of computing the prior distributions (not shown) may also be included in an alternative implementation.

Another receiving operation 204 receives the independent, identically distributed data. Exemplary data may include without limitation auditory speech data from an unknown number of speakers, where determining the correct number of speakers is part of the modeling process and image segmentation data from images containing few large and relatively homogeneous regions as well as several very small regions of different characteristics (outlier regions), where modeling of the few larger regions should not be notably affected by the presence of the outlier regions.

Yet another receiving operation 206 receives initial estimates of the posterior distributions for a set of modeling parameters for a mixture of Student distributions. The initial estimates may be received from another process or be determined in a determining operation (not shown) using a variety of methods, including a random approach. However, the optimization of the modeling parameter can resolve quicker if the initial estimates are closer to the actual posterior distributions. In one implementation, heuristics are applied to the prior distributions to determine these initial estimates. In a simple example, the posteriors are set equal to the priors. A more elaborate example is to heuristically combine the priors with the results of fast, non-probabilistic methods, such as *K*-means clustering.

A selection operation 208 selects one of the modeling parameters in the set of modeling parameters. A computation operation 210 computes a tractable approximation of the posterior distribution of the selected modeling parameter using the current estimates of the other modeling parameters. (In the first iteration, the current estimates of the other modeling parameters represent their initial estimates.) In one implementation, the current state of the estimate of each modeling parameter is stored in a storage location, such as in a memory.

In the illustrated implementation, a variational inference method produces the tractable approximation. In one variational inference approach, the tractable posterior distribution is approximated using the Equations (12), (18), (22), (25), and (28). The tractable approximation of the selected modeling parameter becomes the current estimate of that modeling parameter, which can be used in

subsequent iterations. Alternatively, other approximation methods, including Monte Carlo techniques, may be employed.

A computation operation 212 computes the lower bound of the log marginal likelihood, such as by using Equation (11). If the lower bound is insufficiently optimized according to the computation operation 212, such as by improving by greater than a given threshold or by some other criterion, a decision operation 214 loops processing back to the selection operation 208, which selects another modeling parameter and repeats operation 210 212 and 214 in a subsequent iteration. However, if the lower bound is sufficiently optimized, processing proceeds to a generation operation 216, which generates the probability density of the data based on the mixture of Student distributions characterized by the current estimates of the modeling parameters (e.g., using Equation (4)).

It should be understood that the order of at least some of the operations in the described process may be altered without altering the results. Furthermore, other methods of determining whether the posterior distribution approximations of the modeling parameters are satisfactorily optimized, including testing whether the individual posterior distribution factors (e.g., $q(\mathbf{s})$) change little in each iteration or testing whether the product (e.g., $q(\theta)$) of the posterior distribution factors changes little in each iteration.

FIG. 3 illustrates an exemplary robust Bayesian mixture modeling system 300. Inputs to the system 300 include input data 302, initial estimates of the modeling parameters 304, and prior distributions of the modeling parameters 306.

A modeling parameter selector 308 selects a modeling parameter that is to be approximated in each iteration. A tractable approximation module 310 receives the inputs and the selection of the modeling parameter to generate a tractable approximation of the selected modeling parameter (e.g., based on variational inference or Monte Carlo techniques). In one implementation, the tractable approximation module 301 also maintains a current state of the estimate of each modeling parameter in a storage location, such as in a memory.

Based on the current estimates of the modeling parameters, including the new approximation of the selected modeling parameter, a lower bound optimizer module 312 computes the lower bound of the log marginal likelihood. If the lower bound fails to satisfy an optimization criterion (such as by increasing more than a threshold amount), the lower bound optimizer module 312 triggers the modeling parameter selector module 308 to select another modeling parameter in a next iteration. Otherwise, the current estimates of the modeling parameters are passed to a data model generator 314, which generates a data model 316 including the probability density of the data based on the mixture of Student distributions characterized by the current estimates of the modeling parameters (e.g., using Equation (4))

The exemplary hardware and operating environment of FIG. 4 for implementing the invention includes a general purpose computing device in the form of a computer 20, including a processing unit 21, a system memory 22, and a system bus 23 that operatively couples various system components include the system memory to the processing unit 21. There may be only one or there may be more than one processing unit 21, such that the processor of computer 20

comprises a single central-processing unit (CPU), or a plurality of processing units, commonly referred to as a parallel processing environment. The computer 20 may be a conventional computer, a distributed computer, or any other type of computer; the invention is not so limited.

The system bus 23 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, a switched fabric, point-to-point connections, and a local bus using any of a variety of bus architectures. The system memory may also be referred to as simply the memory, and includes read only memory (ROM) 24 and random access memory (RAM) 25. A basic input/output system (BIOS) 26, containing the basic routines that help to transfer information between elements within the computer 20, such as during start-up, is stored in ROM 24. The computer 20 further includes a hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or writing to a removable optical disk 31 such as a CD ROM or other optical media.

The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive interface 33, and an optical disk drive interface 34, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for the computer 20. It should be appreciated by those skilled in the art that any type of computer-readable media which can store data that is accessible by a computer, such as magnetic cassettes, flash memory cards, digital video disks,

random access memories (RAMs), read only memories (ROMs), and the like, may be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24, or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the personal computer 20 through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port, or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, computers typically include other peripheral output devices (not shown), such as speakers and printers.

The computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as remote computer 49. These logical connections are achieved by a communication device coupled to or a part of the computer 20; the invention is not limited to a particular type of communications device. The remote computer 49 may be another computer, a server, a router, a network PC, a client, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 20, although only a memory storage device 50 has been illustrated in FIG. 4. The logical connections depicted in FIG. 4 include a local-area network

(LAN) 51 and a wide-area network (WAN) 52. Such networking environments are commonplace in office networks, enterprise-wide computer networks, intranets and the Internet, which are all types of networks.

When used in a LAN-networking environment, the computer 20 is connected to the local network 51 through a network interface or adapter 53, which is one type of communications device. When used in a WAN-networking environment, the computer 20 typically includes a modem 54, a network adapter, a type of communications device, or any other type of communications device for establishing communications over the wide area network 52. The modem 54, which may be internal or external, is connected to the system bus 23 via the serial port interface 46. In a networked environment, program modules depicted relative to the personal computer 20, or portions thereof, may be stored in the remote memory storage device. It is appreciated that the network connections shown are exemplary and other means of and communications devices for establishing a communications link between the computers may be used.

In an exemplary implementation, a modeling parameter selector, a tractable approximation module, a lower bound optimizer module, a data model generator, and other modules may be incorporated as part of the operating system 35, application programs 36, or other program modules 37. Initial modeling parameter estimates, input data, modeling parameter priors, and other data may be stored as program data 38.

The embodiments of the invention described herein are implemented as logical steps in one or more computer systems. The logical operations of the present invention are implemented (1) as a sequence of processor-implemented

steps executing in one or more computer systems and (2) as interconnected machine modules within one or more computer systems. The implementation is a matter of choice, dependent on the performance requirements of the computer system implementing the invention. Accordingly, the logical operations making up the embodiments of the invention described herein are referred to variously as operations, steps, objects, or modules.

The above specification, examples and data provide a complete description of the structure and use of exemplary embodiments of the invention. Since many embodiments of the invention can be made without departing from the spirit and scope of the invention, the invention resides in the claims hereinafter appended.